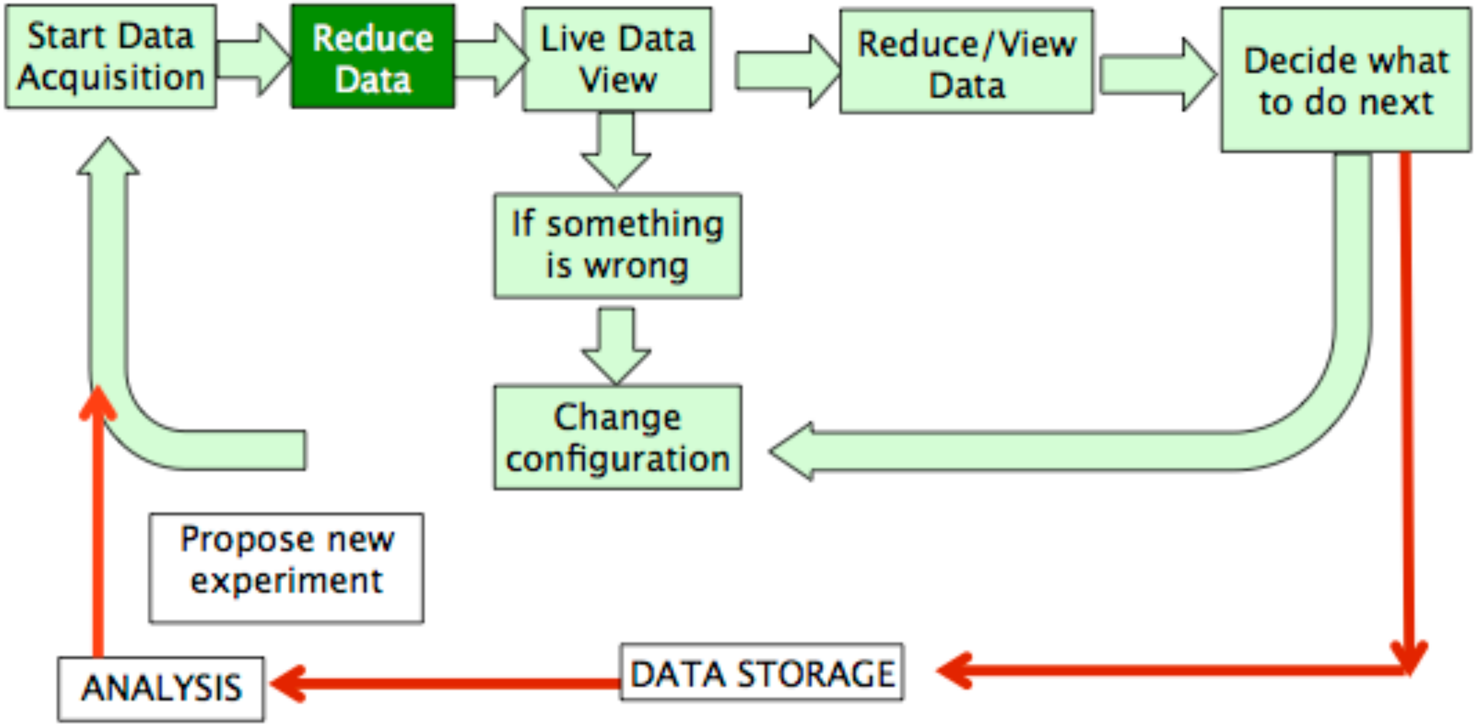
ICEE – Data Mining part (summary from the proposal)

Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

Knowledge discovery & data mining problems

- Discovering and mining expert knowledge from the workflow provenance data
  - Accumulated and collected provenance information
  - Study the viability of predicting data access requirements for the entire set of users.
  - Various clustering or classification algorithms (e.g. model-based clustering, high-dimensional data clustering, pairwise data clustering, support vector machine, etc.) can be used
- Data access pattern recognition
  - Increasing data pre-caching chances to reduce data acquisition latency
  - Keeping up with a quickly evolving set of user requirements
  - The data access patterns can be used to select indexing options for the data and to minimize the data movement in workflow orchestration.

Workflow for near-line analysis and post-processing



From the proposal, figure 2

Simulated experiments

- **Training young scientists without the expense of costly experiments**
  - Running experiments on a Tokamak facility such as KSTAR is expensive, and may be too complicated for inexperienced users.
  - Also, experimental opportunities for inexperienced young scientists are rare.
  - Lack of experiences in young scientists can cost time and efforts in real experiments in production causing faulty experiments with wrong values that may put the facility in jeopardy of destruction from over-heating.
  - Simulations that would replace experiments are expensive to model and build, and simulation capabilities are sometimes not reaching to the experimental capabilities yet.

Training systems

- **Knowledge discovery in a collaborative workflow system based on the workflow provenance data**
  - Enables simulated runs of the experiments based on the previous experimental cases by experts.
  - Workflows and experimental results produce cases that each condition or collection of conditions can be captured as knowledge from the workflow provenance data.
  - The relevant knowledge can be retrieved and used for reasoning new cases for simulated experiments, and adapted and projected for the expected experimental outcomes.
  - The new knowledge can be revised by an expert or through a real experiment, and retained for the next case for simulated experiments.
  - This process based on Case-Based Reasoning (CBR) can help training inexperienced users without the expense of costly experiments, and help gaining valuable access to the simulated facility

Validating workflows

- **Validating workflows for changing conditions based on machine learning**
  - Knowledge discovery in the collaborator workflow system can validate real-time or near-line workflow conditions based on the previous experimental cases.
  - Plan to study adaptive CBR for the dynamic workflow validation model to validate the changing conditions, and provide the workflow system with learning capabilities.
  - This additional validation capability may prevent accidental control values as well as project the workflow results before executing the workflow.

Integrated security

- Dynamic workflow control must be executed in a restricted and privileged mode as well as in a collaborative mode.
- For example, SSL, PKI, OpenID and SAML.
- Plan to leverage the security service design in Earth System Grid Federation (ESGF)
  - where the similar federated authentication and authorization infrastructure needs are present.
- Plan to adopt, for further strict access to the Tokamak workflow control, One-Time-Password (OTP) technology

Security services

- **Authentication Service**
  - responsible for validating user’s authentication information
  - E.g. OpenID in the collaboration.
- **Authorization Service**
  - responsible for validating authorization information based on the user authentication information
  - attaching the user’s access control attributes for a trusted client in the form of a digitally signed SAML statement.
- **One-Time Password (OTP) over Short Messaging Service (SMS)**
  - responsible for delivering OTP over text messaging based on the authorization information for further strict access to the Tokamak workflow control

Mobile access to workflow monitoring

- To explore tablet mobile computing support in collaborative environment, plan to study the following
  - Remote monitoring and control of the workflow and data analysis
    - Mobile distributed access to the monitoring information for the workflow and data analysis, through tablets with Android OS or iOS
    - Tablet access to monitor and update the workflow
    - Tablet access to the results of the workflow
    - Tablet access to analysis results and science discovery
  - Exploration of mobile tablet computing in science
    - Exploring mobile tablet computing as a data collection mechanism for distributed science collaboration
    - Exploring mobile tablet computing as a computing resource for data analysis
    - Exploring mobile tablet computing as a collaborative tool – for example, collaboration in grouping and associating data, analysis, experiments or workflows through tagging and annotation over portable devices